融合双重注意力网络的儿童骨龄评估方法 *

张 鑫，张俊华†，张 帅

(云南大学 信息学院，昆明 650500)

摘 要：骨龄评估是一种检测儿童内分泌与生长发育异常的常用方法，但深度学习方法中低质量手部 X 射线图像降低最终评估精度。针对该问题，提出一种增加手部 X 射线图像感兴趣区域面积的对齐网络，该网络以 Swin Transformer 结构作为主干网络学习图像手部相似性并取得仿射系数，且在训练过程中不需进行大规模手部标注。在骨龄评估网络中，针对高效通道注意力和空间注意力机制改进，提出双池化高效通道注意力和非对称卷积空间注意力方法，将这两种方法以双重注意力形式和 Xception 网络相结合提出 DA-Xception。在 RSNA 数据集上进行测试，该骨龄评估方法达到 5.37 个月的平均绝对误差，相较于其他深度学习方法可更充分提取特征，优化评估结果。

关键词：骨龄评估；X 射线图像对齐；双重注意力；深度学习

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2022.03.0110

Pediatric bone age assessment method combined with dual attention network

Zhang Xin, Zhang Junhua†, Zhang Shuai

(*School of Information Science & Engineering*, *Yunnan University*, *Kunming* 650500, *China*)

Abstract: Bone age assessment is a common method to detect endocrine and growth abnormalities in children. But in deep learning methods, low-quality hand X-ray images reduce the final evaluation accuracy. To solve this problem, this paper proposed an alignment network that increases the area of interest in hand X-ray images. This network uses the Swin Transformer structure as the backbone network to learn image hand similarity and obtain affine coefficients and does not require large-scale hand annotation during the training process. In the bone age assessment network, for the improvement of efficient channel attention and spatial attention mechanism. This paper proposed dual-pool efficient channel attention and asymmetric convolution spatial attention method and combines these two methods in the form of dual attention and Xception network proposes DA-Xception. This paper tested the RSNA dataset, which achieved a mean absolute error of 5.37 months for this bone age assessment method. Compared with other deep learning methods, this method can fully extract features and optimize the evaluation results.

Key words: bone age assessment; X-ray image alignment; dual attention; deep learning

0 引言

人的发育年龄可分为年代学年龄和生物学年龄，其中生物学年龄更加客观地反映了人类实际生长发育情况。通过判断受试者手骨[1]和牙齿生长成熟程度[2]是生物学年龄主要依据。

手骨骨龄评估被广泛应用于现代儿科临床诊断中，医生通过分析受试者非惯用手部 X 射线图像得到对应骨龄，通过与年代年龄进行对比，可判断出儿童生长潜力和骨骼成熟的程度。此外，骨龄评估也为儿童身高发育情况提供参考[3]。

目前，传统的骨龄评估方法有图谱法和计分法，图谱法通过受试者手部 X 射线图像与标准图谱集对比，以最接近标准图谱中图像的标签为受试者骨龄，常见的图谱法有 Greulich-Pyle 图谱法[4]，该方法平均误差为 11.5 个月[5]。计分法通过对手部 X 射线中若干具有代表性的骨骼分别评分，最后计算总分并用公式转换为对应的骨龄，常见的计分法有 TW 计分法[6]和中华 05 计分法[7]。但无论图谱法还是计分法都存在明显弊端，图谱法的结果受评估者主观因素影响导致评估结果误差大；计分法由于要对腕骨、骨骺等手骨区域分别评分，故存在耗时长、效率低的缺陷。

随着计算机视觉技术兴起，自动化骨龄评估方法得以发展，早期的自动评估方法是对人工评估使用的特征进行自动提取。例如，Thodberg 等人[8]设计出 BoneXpert 系统作为骨龄评估商用性软件，该系统通过分析手部中的感兴趣区域 (region of interest, RoI)，并对 RoI 区域评分得到骨龄结果。

近年来，Spampinato 等人[9]采用深度学习方法进行骨龄自动化评估，提出由卷积神经网络搭建的端到端 BoNet 系统，该方法最终误差为 9.5 个月。此后多种基于深度学习的骨龄评估方法被提出[10~12]，其中众多骨龄评估方法于北美放射学会提供的手部 X 射线图像公开数据集上判断性能和精度。2019 年，Wu 等人[13]将手部 X 射线图像先使用 Mask R-CNN 网络进行分割，去除图像中干扰噪声，然后采用注意力残差子网络(residual attention subnet, RAS)进行骨龄回归，该方法最终平均误差为 7.38 个月，然而该方法在图像分割时需要大量工作对手部区域标注，影响骨龄评估效率。Liu 等人[14]提出 VGG-U-Net 网络同样进行手部分割，该网络将 VGG16[15]预训练网络模型代替 U-Net[16]下采样层，改善小样本数据集分割精度，后续利用骨龄标签之间的相互关联性，在数据集上测试结果为 6.05 个月。2020 年，Hao 等人[17]提出 OCNet 骨龄多分类方法代替骨龄回归，并依据手部发育连续性这一特征，通过骨龄评估模型得到三个不同骨龄范围值，最后计算范围重合情况得到骨龄，该方法最终误差为 5.84 个月。2021 年，He 等人[18]首先对手部 X 射线图像进行无损压缩再用 SE-ResNet 网络提取特征

收稿日期：2022-03-10；**修回日期：**2022-05-05 **基金项目：**国家自然科学基金资助项目(62063034，61841112)；云南大学研究生实践创新项目(2021Z50)

作者简介：张鑫(1996-)，男，四川达州人，硕士，主要研究方向为深度学习、生物医学图像处理；张俊华(1976-)，女(通信作者)，云南昆明人，教授，博导，博士，主要研究方向为医学影像处理和分析、模式识别(jhzhang@ynu.edu.cn)；张帅(1997-)，男，湖南衡阳人，硕士，主要研究方向为生物医学图像处理、计算机视觉.

得到骨龄，最后得到 6.04 个月的平均绝对误差，此方法提升了输入图像的手部占比，但未处理图像中存在的干扰信息，图像质量未得到改善。Salim 等人[19]在图像分割基础上，将岭回归层加入骨龄评估模型中而提出岭回归网络(ridge regression network, RRNet)，该方法测试的绝对误差值为 6.38 个月。

从上述方法中得知，手部 X 射线图像质量的好坏影响最终骨龄评估的精度，而对大规模低质量图像分割处理会降低骨龄评估效率。针对这一问题，本文对数据集中图像进行处理，减少图像噪声、对比度不统一情况。针对数据集中不同图像手部 RoI 存在差异，本文引入对齐网络，使数据集中原始图像与标准手部 X 射线图像具有一致的手部结构，减小数据集中手部尺寸和角度不同的影响。为了强化骨龄评估网络特征提取能力，本文在骨龄回归评估中设计双重注意力 Xception 网络(dual attention Xception, DA-Xception)，该网络通过双支路并行学习图像空间和通道中的手部 RoI，最终进行特征融合回归得到骨龄评估结果。

与当前骨龄评估方法相比，本文的主要工作和贡献如下：1)引入手部图像对齐网络，依据图像手部结构的相似性，保证在骨龄评估中手部 RoI 区域的一致，与图像分割相比，采用手部对齐不需要对图像大规模标注，增大骨龄评估中的有效 RoI 面积。2)在骨龄回归网络中设计 DA-Xception 网络，提出双池化高效通道注意力(double pooling efficient channel attention, DPECA)，强化图像通道中整体和纹理特征。3)提出非对称卷积空间注意力(asymmetric convolution spatial attention, ACSA)提取图像空间中细粒度特征，最终在数据集上实验证明该方法优于其他骨龄评估方法。

1　研究方法

本文的骨龄评估方法主要由两个主干网络组成，如图 1 所示。首先是提升原始图像质量和手部位置校准的手部对齐网络，然后是提取手部 RoI 信息后的骨龄回归网络。受近年来 ViT (vision transformer)[20]网络在计算机视觉任务表现优异的启发，本文在对齐网络中引入 Swin Transformer 网络[21]作为主干网络提取特征，然后对图像特征降维接入全连接层，最终得到原始输入图像与标准图像之间的仿射关系，并利用仿射系数使原始图像具有标准手部结构，减少 X 射线图像中手部尺寸和角度差异对最终评估结果的影响。在回归网络中，将尺寸为 299×299 的手部图像输入 DA-Xception 网络提取手部 RoI 特征信息，然后通过全局平均池化层降低通道维度并与输入性别特征编码后的进行融合，后接入 512 个神经元的全连接层和 dropout 层，最终通过单个神经元的全连接层回归得到图像骨龄结果。

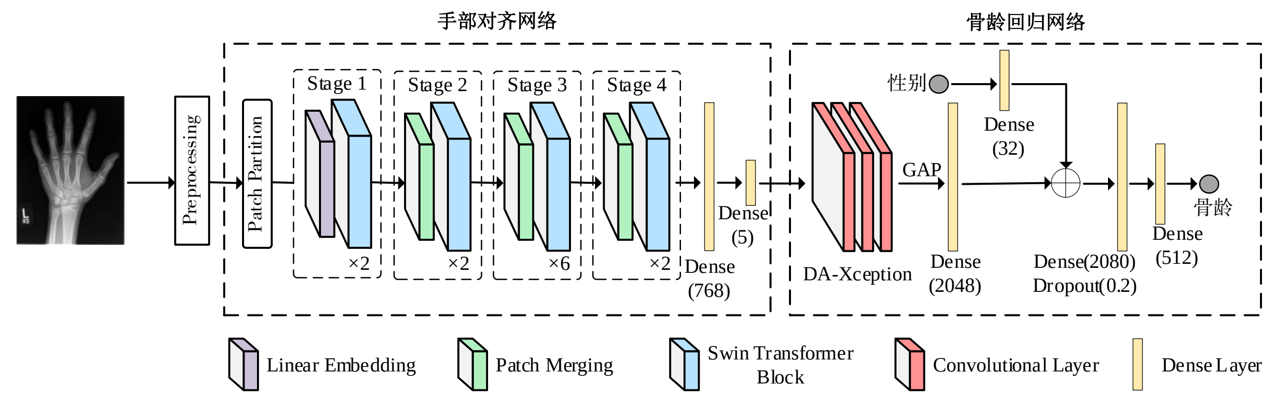


图 1　骨龄评估总体的网络模型图

Fig. 1　The overall network structure of bone age assessment

1.1　手部对齐网络

骨龄数据集中原始 X 射线图像手部通常存在旋转、移动和不同比例的变化，并且这种情况同样存在其他 X 射线数据集[22]。针对上述问题本文引入手部对齐网络，以减少原始图像中手部 X 射线形态变化情况，增加后续骨龄回归中的特征提取能力，对齐网络工作过程如图 2 所示。首先原始图像输入 Swin Transformer 网络中提取手部特征进行训练，并以标准图像为标签，输出图像仿射系数，原始的输入图像通过仿射变换对齐到标准图像。

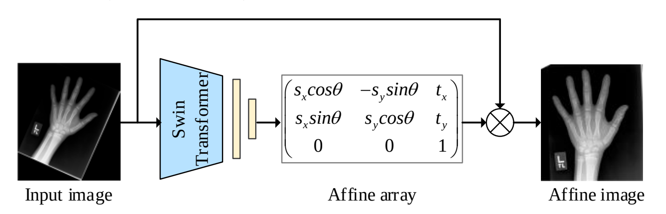


图 2　手部 X 射线图像的对齐原理

Fig. 2　Principles of alignment of hand X-ray images

对于给定的输入图像 I 和标准图像 T，对齐网络通过得到仿射系数 φ，使得仿射图像 $\varphi(I)$具有与标准图像相似的标准手部结构，在网络训练过程中结构损失可定义为 $L_S=f(\varphi(I),T)$。

手部 X 射线图像输入对齐网络得到由五个参数组成的仿射系数 φ，具体形式为 $\varphi=(t_x, t_y, s_x, s_y, \theta)$。其中，$t_x$ 和 t_y 表示图像在水平和垂直方向上的位移程度，s_x 和 s_y 表示水平与垂直方向上的缩放大小，θ 为旋转角度。原始图像经过仿射系数转换的仿射关系如式(1)所示。

$$\varphi(I)=B\left(\begin{pmatrix} s_x cos\theta & -s_y sin\theta & t_x \\ s_x sin\theta & s_y cos\theta & t_y \\ 0 & 0 & 1 \end{pmatrix} G(I), I\right) \tag{1}$$

其中 G 为图像栅格化，B 为图像双线性插值，两者可减少图像在仿射变换过程中特征信息的丢失。

卷积神经网络(Convolutional Neural Network, CNN)分层提取图像特征，通过局部连接(Local Connectivity)和参数共享(Parameter Sharing)方式增强特征提取能力、减少网络参数量，但对图像全局信息的关注度不足。而 ViT 网络在自然语言处理(Natural Language Processing, NLP)中引入多头注意力弥补这一缺陷，但由于自然语言和图像输入规模存在差异，使 ViT 网络计算复杂度大。而 Swin Transformer 网络分层选用不同的采样值减小计算量，并提出滑动窗口方法允许局部特征跨窗口连接提高效率，增大网络感受野。

本文的 Swin Transformer 主干网络中可分为 4 个阶段(Stage)，每个阶段分别由 2，2，6，2 个 Swin Transformer Block 构成，图 3 为连续的两个基本模块，其中包含滑动窗口的多头自注意力机制(Multi-head Self Attention, MSA)、多层感知机(Multilayer Perceptron, MLP)、归一化层(LayerNorm, LN)。

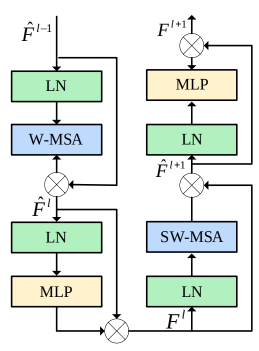


图 3　连续两个 Swin Transformer 基本单元

Fig. 3　Two successive Swin Transformer blocks

本文引入余弦相似度作为对齐网络的结构损失，图像之间的余弦相似度将两张图像转换为向量，并通过计算向量之间夹角的余弦值作为图像相似度关系。故对齐网络损失函数如式(2)所示。

$$L_s = 1 - \cos < I, T > = \frac{\|I\|_2 \times \|T\|_2 - \sum_{i=1}^{n} I_i \times T_i}{\|I\|_2 \times \|T\|_2} \quad (2)$$

其中，I 为原始图像向量，T 为标准图像向量。

1.2　骨龄回归网络

本文的骨龄回归网络为 DA-Xception，该网络基于 Xception 卷积神经网络[23]和双重注意力机制结合而提出，原始 Xception 结构如图 4 所示。

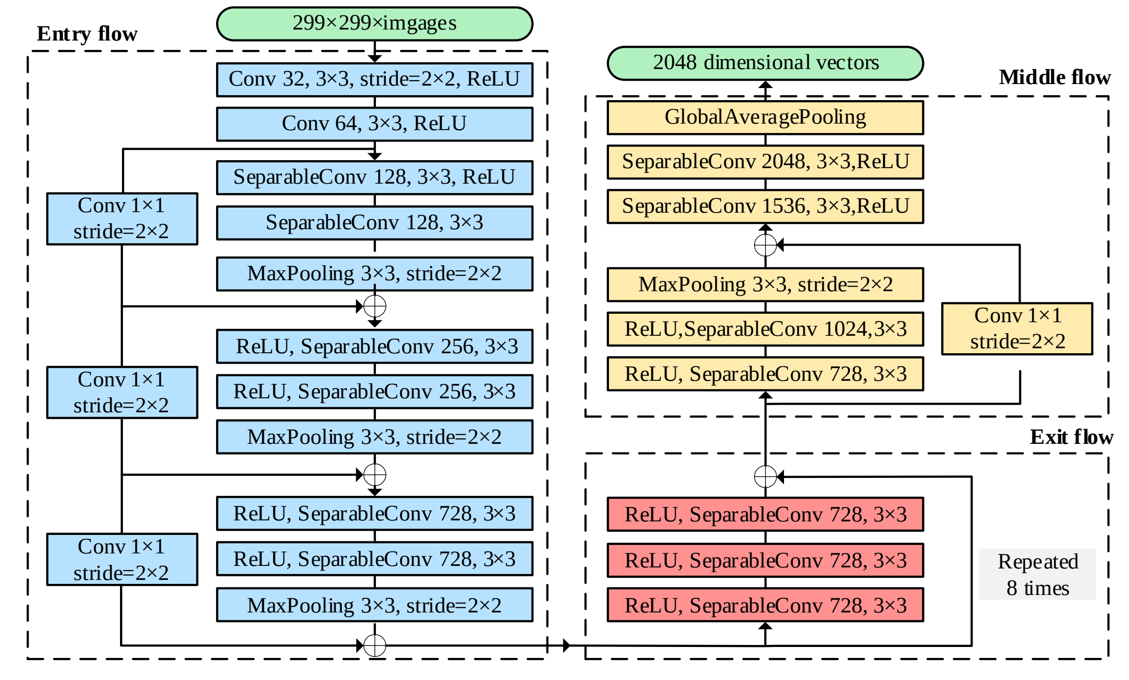


图 4　原始 Xception 网络结构

Fig. 4　Original Xception network structure

1.2.1 双重注意力 Xception

Xception 原始网络模型由多个深度可分离卷积层组成，除去网络的首尾部分，其余卷积模块均采用线性残差方式连接，根据数据输入 Xception 网络先后处理顺序，将图 4 网络结构主要分为三个模块，分别为入口流(entry flow)、中间流(middle flow)和出口流(exit flow)。相比于 Inception V3 网络[24]，两者参数量相近，但 Xception 网络采用深度可分离卷积降低网络运算量，该体系结构更有效地使用模型参数，达到更好的特征提取性能。

原始 Xception 结构虽然具有残差连接模块，但总体为线性神经网络，网络模型只能按照顺序提取特征，这种方式无法充分提取特征中通道与空间信息。Fu 等人[25]为强化语义分割中特征相关联性提出 Dual Attention Network (DA-Net)，该网络采用并行的空间和通道注意力结构同时提取图像通道和位置上的特征。Lin 等人[26]提出双线性(Bilinear)模型进行图像分类，将网络内部分为两个分支结构，并利用不同分支并行提取图像中特征，最后采用双线性池化操作将两个分支特征融合，输出分类结果。Liu 等人[27]针对遥感图像的目标检测提出 Center Boundary Dual Attention Network(CBDA-Net)，该网络以双重结构方式生成中心区域注意力和边界区域注意力，最终在目标检测时能够消除背景噪声干扰并检测到关键区域。

受上述工作启发并结合 Xception 网络结构特性，本文设计出如图 5 所示 DA-Xception 网络。

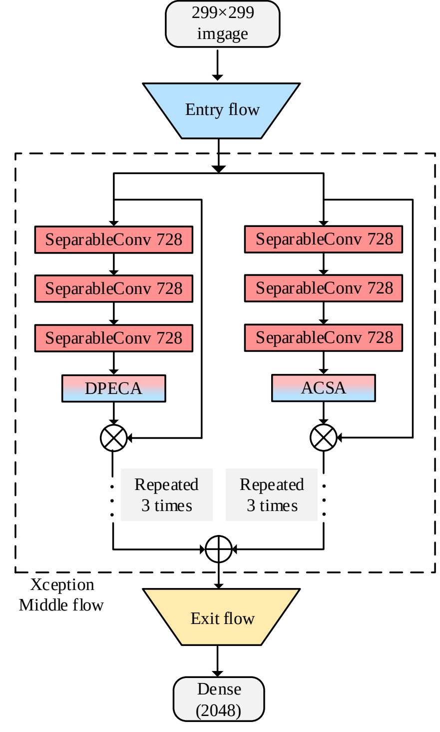


图 5　DA-Xception 网络结构

Fig. 5　Dual attention Xception network structure

相比于原始 Xception 网络中 Middle flow 结构重复 8 次线性残差模块(Residual block)连接，DA-Xception 网络将 Middle flow 模块以双线性结构左右两个分支分别重复 4 次，并在每条支路的残差块中多次加入 DPECA 模块与 ACSA 模块，在左支路中，采用 DPECA 加强图像通道之间特征信息，右支路则对图像空间相关联特征学习。Middle flow 中的双线性结构并行对通道和空间特征信息提取，两条支路上的特征在 Exit flow 结构中特征融合，然后接入全连接层得到骨龄评估结果。

相比于 DA-Net 网络中在输出结构之前单次加入双重注意力，本文所提出的 DA-Xception 网络结构在 Middle flow 的双分支上多次加入通道和空间注意力机制，引导两条支路分别提取图像中通道与空间特征。双线性模型[26]利用双分支学习特征，但左右两条支路结构一致，使最终特征融合时可能存在冗余。本文所提出的双线性融合方法在进行骨龄评估时，有效的提取到图像手部特征，消除了背景噪声干扰。

1.2.2 双池化高效通道注意力

注意力机制通过增加训练过程中感兴趣区域特征张量权重，降低无意义背景的权重，以提高网络特征提取能力。Wang[28]提出图 6(a)所示高效通道注意力(Efficient Channel Attention, ECA)模块在关注通道间注意力中避免降低通道维度，同时以轻量级的方式捕捉跨信道交互。

但 ECA 中仅对输入特征进行全局平均池化(Global Average Pooling, GAP)仅提取图像整体背景信息，却丢失了图像中纹理特征提取。在此基础上本文提出如图 6(b)所示 DPECA 结构，对输入特征进行 GAP 与全局最大池化(Global Maximum Pooling, GAP)操作，强化骨龄评估网络对图像中手部整体与纹理特征充分提取。

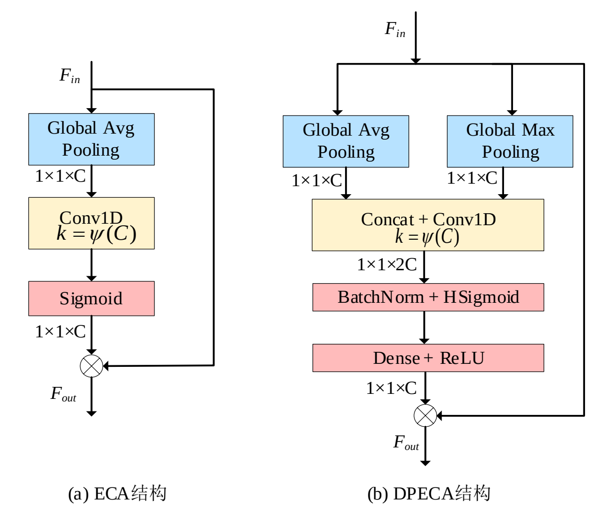


图 6 注意力结构改进对比

Fig. 6 Attention structure improvement comparison

对于 DPECA 输入特征 $F_{in}\in R^{W\times H\times C}$，通过 GAP 与 GMP 后的图像特征定义为 F_{gap} 和 F_{gmp}，其对应公式为式(3)和式(4)。

$$F_{gap}=\max(F_{in}^{c}(i,j)) \quad 0<i<H, 0<j<W \tag{3}$$

$$F_{gmp}=\frac{1}{H\times W}\sum_{u=1}^{H}\sum_{j=1}^{W}F_{in}^{c}(i,j) \quad 0<i<H, 0<j<W \tag{4}$$

其中 $max(\bullet)$表示取特征图像对应通道中的最大值，W、H 和 C 分别表示输入特征的长度、高度和通道个数。池化后的特征进行连接并通过相同一维卷积层共享权重参数，如式(5)所示。最后两个特征经过全连接层聚合生成通道间的注意力关系。

$$M_F=\sigma(\mathrm{Con1D}_k([F_{gap},F_{gmp}])) \tag{5}$$

其中 Con1D 表示一维卷积操作。σ 为 HSigmoid 非线性激活函数。HSigmoid 函数采用分段拟合方式实现 Sigmoid 函数，但 HSigmoid 在网络训练时计算速度更快，具体公式如式(6)所示。

$$f(x)=\begin{cases}0, & (x\le -3)\\ \frac{x}{6}+0.5, & (-3\le x\le 3)\\ 1, & (x\le 3)\end{cases} \tag{6}$$

为了实现合适的跨通道交互作用，卷积核尺寸 k 通过特征图的通道 C 数量自适应进行选择，两者对应关系如式(7)所示。

$$k=\psi(C)=\left|\frac{log_2^{(C)}}{\gamma}+\frac{b}{\gamma}\right|_{odd} \tag{7}$$

其中，$|\bullet|_{odd}$ 表示取最接近运算结果的奇数。

1.2.3 非对称卷积空间注意力

Sanghyun 提出卷积注意力模块(Convolutional Block Attention Module, CBAM) [29]，本文基于 CBAM 结构中空间注意力(Spatial Attention, SA)提出如图 7 所示的 ACSA 结构。

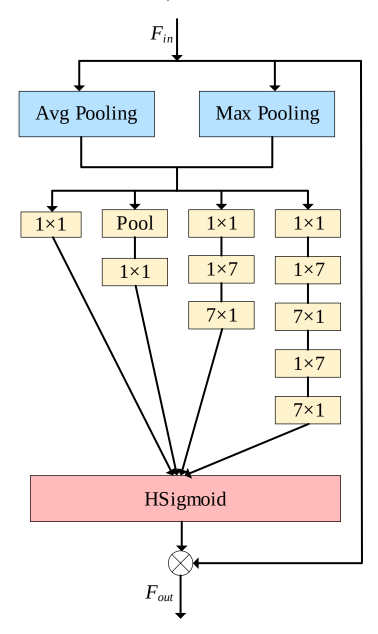


图 7 非对称卷积空间注意力机制结构

Fig. 7 Asymmetric convolution spatial attention module structure

在 ACSA 中，先对图像特征分别平均池化和最大池化操作聚合图像的空间信息，生成两个权重矩阵分别为 $F_{avg}\in R^{W\times H\times 1}$ 和 $F_{max}\in R^{W\times H\times 1}$，然后采用 1×7 卷积核和 7×1 卷积核组代替 CBAM 中 7×7 的卷积核，保证卷积核的感受野不变，提取图像空间中细粒度信息，减小计算量，将两种权重结合得到空间注意力 $M_s(F)$，具体计算如式(8)所示。

$$M_s(F)=\sigma(f^{7\times 7}([AvgPool(F);MaxPool(F)])) \tag{8}$$

其中 $f^{7\times 7}$ 为非对称卷积核组。该空间注意力机制加权选择性地聚合空间特征相似程度，不同空间位置特征，注意力权重由两个位置的特征相似度决定，相似特征表现出更相关联的特性。

1.3 模型损失函数与评价指标

骨龄评估作为一项回归任务，评估模型的最终输出结果为具体实数值。因此，本文选取均方根误差(Root Mean Square Error, RMSE)作为损失函数，其计算公式如式(9)所示。RMSE 损失函数相对于平均绝对误差(Mean Absolute Error, MAE)损失在回归中表现出非线性的损失值降低，在损失较大时，网络模型梯度下降快，使网络快速收敛。损失较小时 RMSE 与 MAE 的值接近，网络模型以线性损失降低。

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{9}$$

其中，N 为样本个数，$\hat{y}_i$ 为模型预测骨龄结果，y_i 为对应标注真实值，从式(9)中可知，随着 RMSE 值减小，模型进行评估结果的优化。

在骨龄评估中，采用平均绝对误差作为指标如式(10)所示。

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

2 实验结果与分析

2.1 数据集

本文的骨龄评估数据集取自 2017 年北美放射学会(Radiological Society of North America, RSNA)举办的儿童骨龄挑战比赛公开数据集。数据集中共有 12811 张图像，包含 6933 张男性和 5878 张女性手骨 X 射线图像，每张图像对应骨骼年龄按月为最小精度进行划分，RSNA 数据集骨龄分布如图 8 所示。

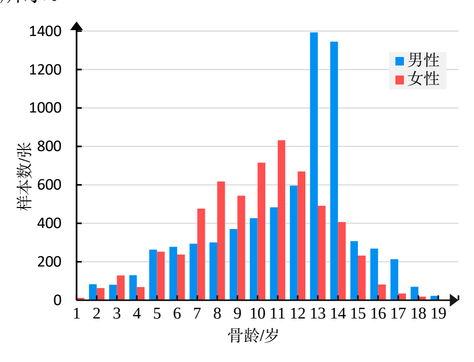


图 8 RSNA 数据集中骨骼年龄分布

Fig. 8 Bone age distribution in RSNA dataset

在骨龄评估结果优化过程中，训练集用于模型的权重更新优化，验证集用于监控模型训练过程并反馈实时训练性能，测试集则对模型的泛化能力进行最终评估。本文实验中随机挑选数据集中 800 张图像进行验证，200 张图像进行测试，其余图像进行网络模型训练。

2.2 实验配置

本文网络模型训练均在 Intel(R) Core(TM) i7-10700KF CPU、NVIDIA GeForce RTX 3070 GPU 和 4×8G 内存的硬件环境下完成，且以 TensorFlow 2.5.0 为深度学习框架、Keras2.6.0 API 的软件环境进行模型训练，编程语言为 Python3.8.12。本文实验中使用 Adam[30]优化器，并将初始学习率设置为 0.001，模型训练进行 100 次迭代，并在训练过程中对验证损失进行监控，在多次未出现损失优化则减小学习率。批量输入大小设置为 16，网络输入图像尺寸为 299×299，采用均方根误差作为网络模型损失，平均绝对误差为评价指标，以衡量骨龄评估值和真实值之间的差距。

2.3 图像预处理

手部 X 射线图像由于采集设备和拍摄曝光方法存在差异，不同图像分辨率和灰度分布不均匀，导致评估结果误差较大。本文对 X 射线图像进行对比度统一和降噪处理，先将 X 射线图像进行直方图均衡化，使图像灰度分布调整到适当范围内，在不影响整体对比度的情况下增强局部对比度；后对 X 射线图像进行自适应伽玛变换使其对比度拉伸，增加图像中低亮度像素值，并抑制高亮度像素减少情况。为了消除图像中的噪声干扰，采用双边滤波进行平滑处理，过程如图 9 所示。

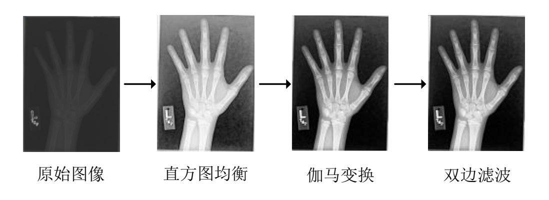


图 9 X 射线图像的预处理过程

Fig. 9 The preprocessing steps of X-ray images

2.4 图像对齐结果

数据集图像经过对齐网络后进行仿射得到标准图像如图 10 所示。图 10(a)中原始图像中手部位置和尺寸存在差异，这些差异使每张图像中的手部 RoI 区域不一致，图 10(b)为手部对齐后的手部 X 射线图像。

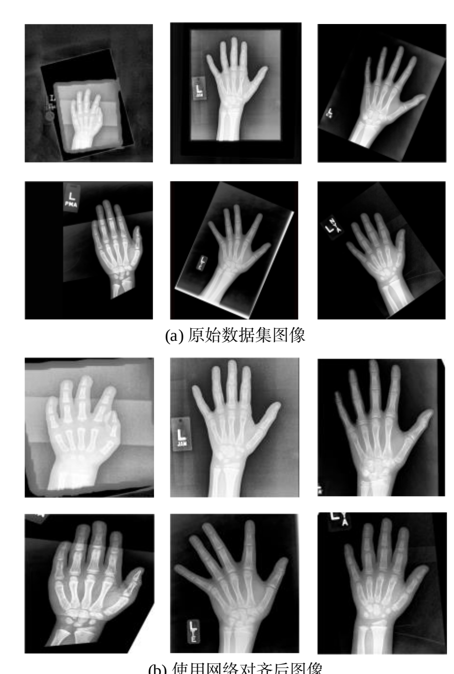


图 10 数据集图像对齐前后对比

Fig. 10 Dataset image alignment before and after comparison

可以看到，对齐后的图像调整了手部倾斜角度和图像中手部区域比例，使手部中感兴趣区域更明显且区域一致，减少后续骨龄回归中错误信息干扰，使有效的手部特征更能够充分地被提取。

2.5 基准网络对比

为了选择合适的网络结构进行骨龄评估工作，本文挑选 EfficientNetB4、ResNet101、DensNet201、Inception ResNet V2 和 Xception 五种经典网络结构进行骨龄评估，并观察评价指标。基准网络评估中对网络结构和数据集不做任何处理，将图像尺寸统一至 299×299 后输入五类网络中得到表 1 所示的五种网络骨龄评估结果。采用原始 Xception 网络评估结果在五组网络中表现最佳，骨龄评估平均绝对误差为 7.41 个月，且该网络参数量适中，因此后续的骨龄评估工作选择 Xception 网络结构进行改进，以此优化最终回归精度。

表 1 不同基准网络的误差与参数量

Tab. 1 MAE and parameter with different baseline networks

网络模型	MAE/月	参数量/107
ResNet101	8.70	4.37
DenseNet201	8.47	1.93
EfficientNetB4	7.90	1.86
Inception ResNet V2	7.46	5.51
Xception	7.41	2.19

2.6 消融与性别实验

为了验证本文骨龄评估方法的有效性，本文对评估过程中的各个模块进行消融实验，衡量不同结构在骨龄评估中的作用。首先验证网络结构中不同模块的有效性；然后探究本文的双重注意力方法与其他注意力改进机制分析对比；最后探讨性别因素对骨龄回归结果影响实验。

2.6.1 不同模块消融实验结果

本文的骨龄评估工作主要分为三个部分，即图像预处理、Swin Transformer 网络将图像对齐和 DA-Xception 网络进行

骨龄回归。对上述三个模块进行消融实验对比，对比实验方法分为：1)仅使用 Xception 网络进行骨龄回归；2)加入图像预处理工作；3)加入图像对齐工作；4)在回归网络中加入 DPECA 和 ACSA 模块，各实验骨龄评估精度如表 2 所示。

表 2　消融实验的 MAE 值

Tab. 2　MAE of ablation experiments

Xception	预处理	图像对齐	双重注意力	MAE/月
√				7.41
√	√			6.67
√	√	√		5.72
√	√	√	√	5.37

在表 2 中，对图像预处理后骨龄评估平均绝对误差值为 6.67 个月，进一步将图像对齐后，骨龄评估误差减小到 5.72 个月，最后将对齐图像输入到 Xception 网络和双重注意力结合的 DA-Xception 网络结构中进行骨龄评估，得到最终误差结果为 5.37 个月，上述三个模块分别使误差结果降低了 0.74 个月、0.95 个月和 0.35 个月。因此在骨龄评估过程中，对图像预处理有效减少了原始图像中存在的噪声，将不同角度和大小的手部图像对齐，增大了图像中有效手部区域占比，保证了手部 RoI 保持一致。将双重注意力加入到 Xception 网络结构中使网络关注到图像中更加丰富的关键特征，提高最终骨龄评估精度。

神经网络的迭代情况反映模型性能好坏，本文选取骨龄评估误差为 5.37 个月的最优模型训练过程曲线如图 11 所示，蓝色曲线表示训练集 MAE 值，红色表示验证集 MAE 值。可以看到，随着迭代次数的增加，训练集与验证集的平均绝对误差值不断减小，并在 20 个轮次后，两者的减小速度变缓并逐渐趋于稳定，当训练 100 轮次时，训练集损失曲线衰减缓慢，此时网络已经充分提取有效特征，骨龄回归结果趋于稳定。

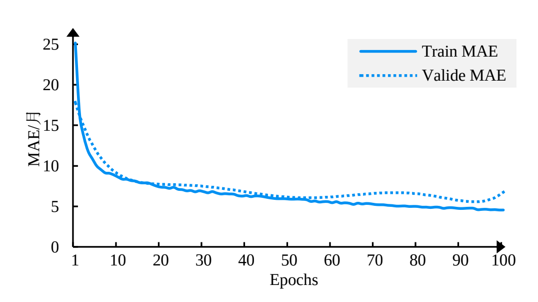


图 11　网络迭代中训练与验证 MAE 曲线

Fig. 11　MAE curve of training and validation

2.6.2 双重注意力有效性验证实验

为了验证本文所提出双重注意力机制的有效性，本小节将本文方法与其他注意力改进机制实验分析对比如表 3 所示。

表 3　不同注意力机制方法结果

Tab. 3　Results of different attention mechanism methods

Methods	MAE/月
Xception	5.72
Xception + Bilinear	5.86
Xception + DA-Net	6.14
Xception + ECA + SA	5.65
Xception + DPECA + ACSA(Ours)	5.37

第一组实验将图像预处理后再对齐，最后采用 Xception 主干网络进行骨龄评估，平均绝对误差为 5.72 个月。第二组在此基础上将 Xception 网络 Middle flow 模块更换为双线性结构且不加入任何注意力机制，其评估误差增加 0.14 个月。第三组将 DA-Net 结构与 Xception 网络融合，得到误差结果为 6.14 个月。第四组将 ECA 模块和 SA 模块以双支路方式与 Xception 网络融合，得到回归误差相对于第三组实验降低 0.49 个月。第五组实验为本文所提出方法，将第四组两个注意模块更换为 DPECA 模块与 ACSA 模块构成双重注意力机制，最终骨龄评估结果 5.37 个月。

为了探究不同网络的骨龄回归方式，本文将后四组实验的网络结构绘制热力图进行网络的可视化，其结果如图 12 所示。每一列对应不同网络结构的注意力图。

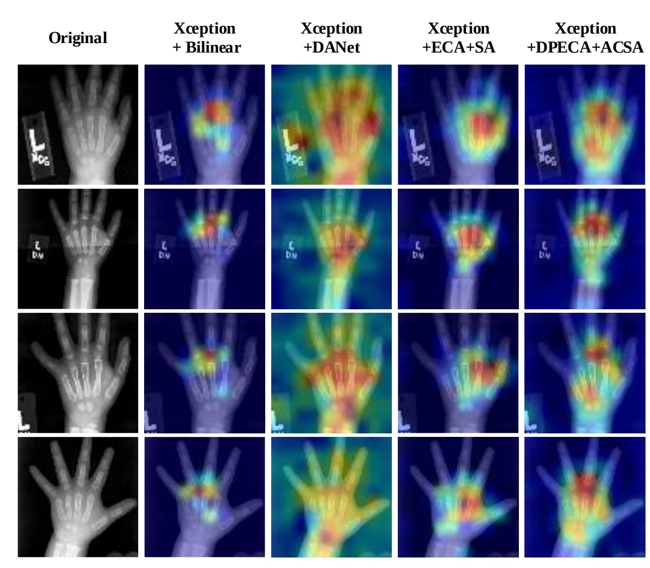


图 12　不同网络结构热力图

Fig. 12　Heatmap of different network structures

可以看到，第二组实验中双线性结构对骨龄评估手部区域关注度不足，因此相对与 Xception 原始网络结构骨龄评估误差增大，第三组实验为 Xception 网络结合 DA-Net，其热力图可以看到该网络中不仅提取手部区域特征，而且关注手部边沿背景信息，网络图像背景特征降低骨龄评估结果。

第四组与第五组实验为在 Xception 网络结构中加入原始的双重注意力与本文所改进双重注意力对比，在图 12 中两种网络结构关注区域相似，但第五组关注手部区域面积更大且更加关注手部关键性区域，从最终骨龄回归结果得知，本文所提出方法在此基础上误差减小 0.28 个月。因此采用 DPECA 与 ACSA 模块，能够更加有效的提取手部特征。

2.6.3 性别因素对比实验

在人的生长发育中，男性与女性在同一年龄段手部发育成熟程度存在差异，探究性别因素对骨龄评估结果如图 13 所示。

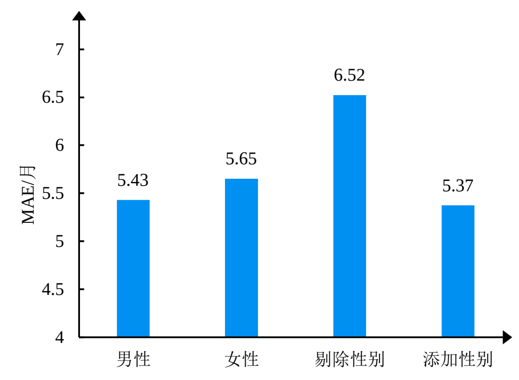


图 13　性别因素对骨龄评估结果影响

Fig. 13　Gender influence in bone age assessment

实验分为四个部分：1)单独对 RSNA 数据集中男性手部 X 射线图像进行骨龄评估；2) 单独对女性进行骨龄评估；3) 在 RSNA 数据集中剔除性别信息进行骨龄评估；4)加入性别信息进行骨龄评估。单独对男性和女性分别进行骨龄评估平均绝对误差为 5.43 和 5.65 个月，不加入性别特征骨龄评估误差为 6.52 个月，在骨龄评估中加入性别信息 MAE 为 5.37

个月。在对单一性别进行骨龄评估时，相比于不加入性别误差分别减小了 1.09 和 0.87 个月，而性别因素的加入使误差减小 1.15 个月，故在骨龄评估中加入性别信息能够有效减小误差值，提高回归精度。

2.7　不同深度学习方法对比分析

为了更好地说明本文方法在骨龄评估中的先进性，本文对近年来其他骨龄评估方法进行比较，表 4 展示了不同方法下的骨龄评估平均绝对误差值。

表 4　不同骨龄评估方法结果

Tab. 4　Results of different bone age assessment methods

Methods	Date	MAE/月
VGG16[12]	2020	9.97
RAS[13]	2019	7.38
RRNet[19]	2021	6.38
Ranking CNN[14]	2019	6.05
SE-ResNet[18]	2021	6.04
OCNet[17]	2020	5.84
DA-Xception(ours)	/	5.37

在表 3 中，文献[12]将骨龄评估工作分为两个阶段，先对图像进行分割，去除手部 X 射线图像中的背景干扰信息，然后采用 VGG16 网络进行骨龄回归，最终回归误差为 9.97 个月。文献[12~14, 19]同样采取上述思路，先对数据集中图像分割手部区域，剔除图像中干扰标签信息，然后将分割图像输入骨龄回归网络最终得到每张图像的骨龄评估结果，这三种方法最终得到回归误差分别为 7.38 个月、6.05 个月和 6.39 个月。图像分割方法有效的去除了图像背景干扰，优化骨龄评估结果，但分割方法需对数据集中图像手部区域进行大量人工标注工作。与上述分割方法相比，本文所提出的将图像手部对齐方法仅需要对少量图像进行弱注释，利用对齐网络自动提取手部 RoI 特征，使每张图像手部的 RoI 区域趋于一致，同时本文方法中图像预处理工作能够抑制图像中的噪声干扰。

文献[18]提出在骨龄评估网络中加入图像的无损压缩模块，在降低图像尺寸时保证图像的质量稳定，随后将图像输入到 SE 注意力机制和 ResNet 网络结合的骨龄回归网络中，最终骨龄评估误差为 6.04 个月。文献[13]加入残差注意力使网络更关注 Roi 区域，相比于在网络中单一加入注意力机制，本文采用双支路方式加入通道与空间并行注意力机制，最终得到 5.37 个月的骨龄评估误差，优于表 3 中的其他方法，进一步提高评估精度。

综上所述，本文方法在骨龄评估中仅需轻量级图像标注信息即可完成图像的对齐工作，不需要过多的人工处理，临床应用可行性更高。此外，采用双重注意力结构进行骨龄回归工作相对线性网络中加入注意力机制能够更充分提取手部特征信息，进一步提高准确性，减小骨龄评估误差。

3　结束语

针对目前骨龄评估中手部 X 射线存在图像质量不高，手部区域在图像中尺寸、角度存在差异情况。本文在改善 X 射线图像质量的基础上，对图像手部进行对齐，同时创新性地提出了将 Xception 网络和两种注意机制通过双支路并行结构方式结合的网络 DA-Xception。

本文的骨龄评估方法分为两个部分，第一部分对 X 射线图像预处理，使图像的对比度、亮度统一，并使用 Swin Transformer 网络进行提取特征使 X 射线图像手部感兴趣区域对齐。第二部分采用 DA-Xception 网络提取手部感兴趣区域特征后骨龄回归得到评估结果。最后通过实验证明本文方法可以有效地减小图像质量对评估结果的影响，并与当前其他骨龄评估方法相比评估结果精度更高，为后续骨龄评估工作提供重要参考价值，更有助于预防青少年生长发育疾病。

参考文献：

[1] Fishman L S. Radiographic evaluation of skeletal maturation: a clinically oriented method based on hand-wrist films [J] The Angle Orthodontist, 1982, 52 (2): 88-112.

[2] Liversidge H M, Molleson T I. Developing permanent tooth length as an estimate of age [J]. Journal of Forensic Science, 1999, 44 (5): 917-920.

[3] Martin D D, Wit J M, Hochberg Z, *et al.* The use of bone age in clinical practice-part 1 [J]. Hormone Research in Paediatrics, 2011, 76 (1): 1-9.

[4] Bayer L M. Radiographic atlas of skeletal development of the hand and wrist [J]. California Medicine, 1959, 91 (1): 53.

[5] King D G, Steventon D M, Osullivan M P, *et al.* Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods [J]. The British Journal of Radiology, 1994, 67 (801): 848-851.

[6] Tanner J M, Oshman D, Bahhage F, *et al.* Tanner-Whitehouse bone age reference values for North American children [J]. Journal of Pediatrics, 1997, 131 (1): 34-40.

[7] 张绍岩，刘丽娟，吴真列，等. 中国人手腕骨发育标准-中华 05L. TW3-CRUS, TW3-C 腕骨和 RUSCHN 方法 [J]. 中国运动医学杂志, 2006, 25 (5): 509-516. (Zhang Shaoyan, Liu Lijuan, Wu Zhenlie, *et al.* The skeletal development standards of hand and wrist for Chinese Children-China 05 I. TW3-CRUS, TW3-C Carpal, and RUS-CHN Methods [J]. Chinese Journal of Sports Medicine, 2006, 25 (5): 509-516.)

[8] Thodberg H H, Kreibor S, Juul A, *et al.* The BoneXpert method for automated determination of skeletal maturity [J]. IEEE Trans on Medical Imaging, 2009, 28 (1): 52-66.

[9] Spampinato C, Palazzo S, Giordano D, *et al.* Deep learning for automated skeletal bone age assessment in X-ray images [J]. Medical Image Analysis, 2017, 36: 41-51.

[10] Igiovikov V I, Rakhlin A, Kalinin A A, *et al.* Paediatric bone age assessment using deep convolutional neural networks [M]. Cham: Springer, 2018: 300-308.

[11] Liang Baoyu, Zhai Yunkai, Tong Chao, *et al.* A deep automated skeletal bone age assessment model via region based convolutional neural network [J]. Future Generation Computer Systems, 2019, 98 (9): 54-59.

[12] Gao Yunyuan, Zhu Tao, Xu Xiaohua. Bone age assessment based on deep convolution neural network incorporated with segmentation [J]. International Journal of Computer Assisted Radiology and Surgery, 2020, 15 (12): 1951-1962.

[13] Wu, Eric, Kong Bin, Wang Xin, *et al.* Residual attention based network for hand bone age assessment [C]// Proc of the 16th International Symposium on Biomedical Imaging. New York: IEEE Access, 2019: 1158-1161.

[14] Liu Bo, Zhang Yu, Chu Meicheng, *et al.* Bone age assessment based on Rank-Monotonicity enhanced ranking CNN [J]. IEEE Access, 2019, 7: 120976-120983.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) . https://arxiv.org/pdf/1409. 1556. pdf.

[16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[17] Hao Pengyi, Xie Xuhang, Han Tianxing, *et al.* Overlap classification mechanism for skeletal bone age assessment [C]// Proc of the 2nd ACM

International Conference on Multimedia in Asia. New York: Association for Computing Machinery, 2021: 1-7.
[18] He Jin, Jiang Dan. Fully automatic model based on SE-ResNet for bone age assessment [J]. IEEE Access, 2021, 9: 62460-62466.
[19] Salim I, Hamza A B. Ridge regression neural network for pediatric bone age assessment [J]. Multimedia Tools and Applications, 2021: 1-18.
[20] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2021-06-03) . https://arxiv. org/pdf/2010. 11929. pdf.
[21] Liu Ze, Lin Yutong, Cao Yue, *et al.* Swin Transformer: hierarchical vision transformer using shifted windows [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 10012-10022.
[22] Liu Jingyu, Zhao Gangming, Fei Yu, *et al.* Align, attend and locate: chest X-Ray diagnosis via contrast induced attention network with limited supervision [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press. 2019: 10632-10641.
[23] Chollet F. Xception: deep learning with depthwise separable convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2017: 1251-1258.
[24] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2016: 2818-2826.
[25] Fu Jun, Liu Jing, Tian Haijie, *et al.* Dual attention network for scene segmentation [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2019: 3146-3154.
[26] Lin T Y, Roy Chowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 1449-1457.
[27] Liu Shuai, Zhang Lu, Lu Huchuan, *et al.* Center-boundary dual attention for oriented object detection in remote sensing images [J]. IEEE Trans on Geoscience and Remote Sensing, 2021, 60: 1-14.
[28] Wang Qilong, Wu Banggu, Zhu Pengfei, *et al.* ECA-net: efficient channel attention for deep convolutional neural net-works [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 11534-11542.
[29] Woo S, Park J, Lee JY, *et al.* CBAM: convolutional block attention module [C]// Proc of the 15th European Conference on Computer Vision. Cham: Springer, 2018: 3-19.
[30] Kingma D, Jimmy B. Adam: a method for stochastic optimization [EB/OL]. (2017-01-30) . https://arxiv. org/pdf/1412. 6980. pdf.